Published online 6 August 2009

Nucleic Acids Research, 2009, Vol. 37, No. 18 ***5943–5958***
doi:10.1093/nar/gkp625

Predicting eukaryotic transcriptional cooperativity by Bayesian network integration of genome-wide data

Yong Wang[1,2], Xiang-Sun Zhang[2] and Yu Xia[1,*]

[1]Bioinformatics Program, Department of Chemistry, Boston University, Boston, MA 02215, USA and
[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China

Received March 12, 2009; Revised and Accepted July 13, 2009

ABSTRACT

Transcriptional cooperativity among several transcription factors (TFs) is believed to be the main mechanism of complexity and precision in transcriptional regulatory programs. Here, we present a Bayesian network framework to reconstruct a high-confidence whole-genome map of transcriptional cooperativity in *Saccharomyces cerevisiae* by integrating a comprehensive list of 15 genomic features. We design a Bayesian network structure to capture the dominant correlations among features and TF cooperativity, and introduce a supervised learning framework with a well-constructed gold-standard dataset. This framework allows us to assess the predictive power of each genomic feature, validate the superior performance of our Bayesian network compared to alternative methods, and integrate genomic features for optimal TF cooperativity prediction. Data integration reveals 159 high-confidence predicted cooperative relationships among 105 TFs, most of which are subsequently validated by literature search. The existing and predicted transcriptional cooperativities can be grouped into three categories based on the combination patterns of the genomic features, providing further biological insights into the different types of TF cooperativity. Our methodology is the first supervised learning approach for predicting transcriptional cooperativity, compares favorably to alternative unsupervised methodologies, and can be applied to other genomic data integration tasks where high-quality gold-standard positive data are scarce.

INTRODUCTION

Transcription factors (TFs) are proteins that dynamically read and interpret the static genetic instructions in the DNA (1,2). TFs usually cooperate with other TFs to facilitate (as an activator) or inhibit (as a repressor) the recruitment of RNA polymerase, using complex logic rules built from simple rules (AND, OR and NOT) to control the precise condition-dependent expression of target genes (TGs) (2–6). Overall, transcriptional cooperativity among several TFs is believed to play an important role in generating complexity and precision in transcriptional regulatory programs, especially in eukaryotic organisms. However, the genome-wide network of eukaryotic TF cooperativity remains largely unknown. Thus research on this topic is fundamentally important and in pressing need.

Although central to transcriptional regulation, transcriptional cooperativity is an intrinsically complex phenomenon (7). In this article, transcriptional cooperativity is broadly defined as the functional interaction between two or more TFs to regulate the expression of a TG. It can refer to TF–TF physical/genetic/regulatory interaction, competitive regulation and TF spatial and temporal combinatorial co-regulation. Furthermore, we focus only on the cooperativity among different DNA-binding TFs, and exclude TF homodimer interactions and interactions between TFs and non-DNA-binding cofactors.

Experimental methods for detecting TF interaction include co-immunoprecipitation and super-gel shift (1). These methods are generally time-consuming, and it is difficult to apply them to mapping the whole-genome TF cooperativity network in the living cell (8–10). Complementarily, a wide variety of computational approaches have been proposed to predict TF cooperativity. Some of them focus on case studies (11–14), whereas others are based on the unsupervised framework using a

*To whom correspondence should be addressed. Tel: 617 358 2302; Fax: 617 353 4814; Email: yuxia@bu.edu

single data source such as TF-binding motif (6,8,15–19), TG (9,20–22) and TF activity (23,24).

A well-known example of TF cooperativity is the TF pair MATa1 and MATα2 that plays an important role in determining the yeast cell type. These two TFs interact cooperatively by forming the heterodimer MATα2/MATa1 which binds DNA with much higher specificity and affinity than each TF alone (25,26). In this case, the existence of TF cooperativity is supported by more than one type of genome-wide data sources. For example, genome-wide protein–protein interaction data shows that MATa1 and MATα2 physically interact with each other; genome-wide transcriptional regulatory data shows that MATa1 and MATα2 share significantly larger number of TGs than expected by chance. In this article, we aim to predict such TF cooperativity by systematically integrating diverse data sources using Bayesian networks. Bayesian networks have been widely used in computational biology, such as in protein structural modeling (27), protein–protein interaction prediction (28), protein function prediction (29–32), gene-expression analysis (33), among others. Here, we apply Bayesian networks for the first time to TF cooperativity prediction. Compared to previous applications, there are many new challenges in applying Bayesian networks to TF cooperativity prediction, such as scarcity of gold-standard data, and assessment of genomic predictors for TF cooperativity. Although there exist many computational approaches to study TF cooperativity (6,8,9,15–24), our method is the first supervised learning approach to integrate genome-wide data sources for predicting TF cooperativity.

We choose *Saccharomyces cerevisiae* as our model eukaryote since many different types of genome-wide data sources are available in yeast. We improve the prediction accuracy of TF cooperativity by combining two strategies. First, we assemble a comprehensive list of genomic data sources, and systematically assess each data source in terms of its predictive power for TF cooperativity. Second, we integrate these heterogeneous data sources to infer TF cooperativity using a supervised learning framework. The key component of our approach is the introduction of a gold-standard positive (GSP) dataset and the construction of Bayesian network for predicting TF cooperativity.

MATERIALS AND METHODS

Gold-standard data collection and feature assessment

We collected 174 TFs from the YEASTRACT database (34,35) and the *Saccharomyces* Genome Database (SGD) (listed in Supplementary Table 1). Last updated in September 2007, the YEASTRACT database contains over 30 980 regulatory associations between TFs and TGs, and includes 284 specific DNA-binding sites for 108 characterized TFs from more than 1000 bibliographic references. The total number of TFs from this database is 170. In addition, we manually add four more TFs annotated in the SGD database.

We compiled 25 TF pairs each belonging to the same biochemically well-defined complex according to the MIPS complex catalogue (36) as our approximate GSP for TF cooperativity (Supplementary Table 2). This choice is reasonable, as co-transcriptional complex relationship is a strong and reliable signal for coordinated actions among TFs. In addition, we constructed an approximate gold-standard negative (GSN) set for TF cooperativity by identifying all TF pairs that do not belong to any known MIPS complex. Our GSP and GSN sets are approximate: the GSP set is the only high-quality dataset of TF cooperativity currently available, and is more restrictive compared to our broad definition of TF cooperativity; at the same time, the GSN set is expected to contain a small fraction of false negatives. Nevertheless, our results in this article suggest that the quality of the GSP and GSN sets are good enough for generating useful predictions of transcriptional cooperativity.

We collected 15 features that potentially correlate with TF cooperativity based on genome-wide information such as sequence, expression, regulation, interaction and function. We converted the numerical features into binary ones by binning the data, and calculated the likelihood ratio scores of every feature in the following way. For each binary feature f taking on a particular value (1 or 0; presence or absence), we count the following four numbers: true positives (TP) which are GSPs where the feature takes on the given value, true negatives (TN) which are GSNs where the feature does not take on the given value, false positives (FP) which are GSNs where the feature takes on the given value, and false negatives (FN) which are GSPs where the feature does not take on the given value. The likelihood ratio is then defined as the fraction of GSPs where the feature takes on the given value, divided by the fraction of GSNs where the feature takes on the given value: $\mathrm{LR} = \mathrm{Prob}(f \mid \mathrm{GSP})/\mathrm{Prob}(f \mid \mathrm{GSN}) = (\mathrm{TP}/\mathrm{GSP})/(\mathrm{FP}/\mathrm{GSN})$. The likelihood ratio scores are central to the Bayesian framework (28,29), and can be regarded as a measure of the usefulness of single features in data integration. A likelihood ratio score much larger than 1 indicates that the feature is a good predictor for TF cooperativity. Likewise, a likelihood ratio score less than 1 indicates that the evidence is anti-predictive for TF cooperativity. Likelihood ratios of different features are directly comparable, and can be multiplied together to obtain the total likelihood ratio of the combined features indicating the confidence level for the combined evidence when the features are conditionally independent (naïve Bayes integration).

TF–TF interaction, co-expression and co-evolution as predictors for TF cooperativity

Known mechanisms of TF cooperativity can be roughly classified as direct cooperativity and indirect cooperativity (5,37). In the case of direct cooperativity, a TF acts through physical interaction with another TF. Existing genome-wide physical interaction data can lend support to this mechanism. In the case of indirect cooperativity, a TF can interact with another TF genetically and jointly produce a phenotype. Existing protein genetic interaction data can lend support to this mechanism (38).

Here, we assembled 150 TF–TF interactions (88 physical interactions and 62 genetic interactions) from BioGRID as a feature for TF cooperativity prediction (Feature F1 in Figure 2). As shown in Figure 1, the existence of TF physical/genetic interaction is strongly predictive for TF cooperativity (likelihood ratio 92.13, Fisher's exact test P-value $<10^{-15}$).

Studies on various model systems have shown that cooperative TFs have complex patterns of co-expression relationships (14), and TF activity and TF cooperativity are condition-dependent (39). Here, a TF pair has co-expression relationship (Feature F2 in Figure 2) if the largest Pearson correlation coefficient (PCC) of their gene-expression profile out of five different conditions (cell cycle, sporulation, pheromone treatment, unfolded protein response and stress response) exceeds 0.7. In other words, a TF pair has co-expression relationship if the two TFs co-express in at least one condition. This strategy captures the dynamic property of TF co-expression relationships compared to directly computing PCC from the assembled expression compendium of five conditions. As shown in Figure 1, the TF co-expression feature is a weak predictor for TF cooperativity (likelihood ratio 1.83, Fisher's exact test P-value <0.05). Even though the likelihood ratio score increases with a more stringent PCC cutoff (Supplementary Figure 2), TF co-expression feature is generally a weak predictor for TF cooperativity with a large number of FPs. This weak predictive performance is not due to the way that the TF co-expression feature is calculated, as our current strategy of choosing the largest PCC of five conditions compares favorably to alternative measures of TF co-expression such as the average (or median) PCC of five conditions, and the overall PCC of expression profile under all five conditions. Rather, it reflects the transient and dynamic nature of TF cooperativity.

The functional relatedness of TFs inferred from comparative genomic data can provide useful information for TF cooperativity. The basic assumption is that a cooperative TF pair tends to co-occur in different genomes, to be close in the chromosome, and to be fused together in another genome. We used the existing comparative genomic data from Prolinks database (40) as the co-evolution feature to predict TF cooperativity (Feature F3 in Figure 2). As shown in Figure 1, TF co-evolution feature is a strong predictor for TF cooperativity (likelihood ratio 81.82, Fisher's exact test P-value $<10^{-5}$).

TF–TG regulatory relationships from ChIP–chip, literature and motif occurrence data

In addition to the direct relationships between TFs in terms of interaction, co-expression and co-evolution, TF cooperativity can also be deduced from the properties of the jointly regulated TGs in the transcriptional regulatory network. This is due to the following two reasons: first, cooperative TFs tend to share significantly larger number of TGs than expected by chance (TG overlap evidence); second, the TGs of cooperative TFs tend to share significant co-expression, co-function and interaction relationships (TG coherence evidence). We follow a two-step process to compute these TG-based features. First, we identify transcriptional regulatory relationships (TF–TG associations) from ChIP–chip, literature and motif occurrence data. Second, we compute TG overlap and TG coherence features based on the TF–TG association datasets.

We compiled three datasets of transcriptional regulatory relationships. The first dataset is based on five global chromatin immunoprecipitation followed by microarray (ChIP–chip) experiments in yeast (37,41–44). This dataset contains 143 TFs, 4705 TGs and 15 814 transcriptional regulations. The second dataset is based on other experiments recorded in the literature, and contains 162 TFs, 4568 TGs and 17 616 transcriptional regulations. The third dataset of TF–TG associations is predicted based on the occurrence of known TF-binding site motifs in the promoter region of the TGs (34,35). We measured the degree of match between a promoter region and a binding site in terms of the number of sites in both DNA strands, and the distance between adjacent sites. We applied a restrictive cutoff by requiring three binding site occurrences within a 200-bp window, and identified 28 142 transcriptional regulations from 281-binding sites and 6712 upstream promoter sequences. Our procedure takes into account factors that are important in determining TF–TG association: frequency of motif occurrence, motif orientation, and inter-motif distance (17,45). Indeed, all three factors contribute to the prediction of TF cooperativity, as measured by the likelihood ratio of the TG overlap score (Supplementary Figures 5 and 6; see the following section for the definition of TG overlap score). The relationships among these three transcriptional regulatory datasets (ChIP–chip, literature and motif occurrence based) are shown as a Venn diagram in Supplementary Figure 3.

It is difficult to integrate these three datasets into a single transcriptional regulatory network due to two reasons. First, different transcriptional regulatory datasets have different qualities. The literature based data are generally better in quality, while ChIP–chip and motif occurrence data are noisy. Motif occurrence data are also incomplete, as only 108 out of a total of 174 TFs have known binding sites. Second, it is difficult to construct a comprehensive and high-quality gold-standard data for transcriptional regulation, which is essential for proper integration of different transcriptional regulation datasets. Instead, in this work we use the three transcriptional regulatory datasets separately to predict TF cooperativity before combining the predictions together in an optimal way.

TG overlap as a predictor for TF cooperativity

For each transcriptional regulatory dataset, we predict TF cooperativity by enumerating all TF pairs where there is a significant overlap of TGs. This is based on the observation that cooperative TFs tend to share more common TGs in the transcriptional regulatory network than expected by chance (22). For a given TF pair, to determine whether the TG overlap is statistically significant, we fix the total number TGs in the yeast genome (N), the number

of TGs regulated by the first TF (N_1), the number of TGs regulated by the second TF (N_2), and treat the number of TGs regulated by both TFs as a random variable X. Under the null hypothesis that the regulation by the first TF is independent of the regulation by the second TF, X follows a hypergeometric distribution:

$$P(X=i)=\frac{\binom{N_1}{i}\binom{N-N_1}{N_2-i}}{\binom{N}{N_2}} \quad \mathbf{1}$$

From here we can then calculate a P-value score, which is defined as the probability that the TG overlap would assume a value greater than or equal to the observed value, m, by chance:

$$P(X\geq m)=1-\sum_{i=0}^{m-1}P(X=i) \quad \mathbf{2}$$

The TG overlap is statistically significant if the P-value score is smaller than a chosen cutoff.

The above procedure gives us all TF pairs for which the TG overlap is statistically significant. To further quantify the extent of the TG overlap, we also calculate an enrichment score, defined as the ratio of the observed TG overlap versus the expected TG overlap by chance, as follows:

$$F=\frac{Nm}{N_1N_2} \quad \mathbf{3}$$

A score larger than 1 indicates that there is more TG overlap than expected by chance.

Both P-value and enrichment scores are predictive for TF cooperativity (Supplementary Figure 4). We further combine the P-value score and the enrichment score into a single TG overlap feature: two TFs share TG overlap if the P-value score and the enrichment score are both more significant than the corresponding pre-defined cutoffs. We use a P-value score cutoff of 10^{-3} for ChIP–chip and literature based TG overlap calculations, and 10^{-4} for motif occurrence based TG overlap calculation. The strict P-value score cutoffs take into account the noisier nature of the motif data, and serve as a correction for multiple hypothesis testing. We use an enrichment score cutoff of 2 for all calculations. In the end, for each TF pair we obtain three TG overlap scores based on ChIP–chip, literature and motif occurrence evidences (Features F4, F5 and F6 in Figure 2). As shown in Figure 1, all three features are strong predictors for TF cooperativity (likelihood ratio >8, Fisher's exact test P-value $<10^{-7}$).

TG coherence as a predictor for TF cooperativity

In addition to computing the overlap of TGs jointly regulated by a TF pair, we can further use the TG coherence information to predict TF cooperativity. Here, TG coherence is defined as the degree of similarity or closeness among TGs jointly regulated by a TF pair, in terms of co-expression, interaction and co-function. The rationale behind this computation is the observation that co-regulated TGs by a cooperative TF pair tend to interact, co-express, or share similar cellular function (20,21,46,47).

To compute TG coherence scores, we first need to quantify co-expression, interaction and co-function for every TG pair. To quantify co-expression for a TG pair, we computed the largest PCC of the gene-expression profile out of five conditions. We then applied the cutoff >0.8 and selected 3.57% of all TG pairs as co-expressed ones. To quantify interaction for a TG pair, we applied the diffusion kernel (48) to the BioGRID protein–protein interaction data (49) to obtain a kernel matrix. We then applied the cutoff >2 to the kernel matrix and selected 0.63% of all TG pairs as interacting ones. This cutoff corresponds roughly to the inclusion of second nearest neighbors in the interaction network. To quantify co-function for a TG pair, we adopted the GO term similarity measure introduced in (50,51), and computed the average GO term similarity by considering all GO biological process terms assigned to the TG pair. We then applied the cutoff >7 and selected 5.85% of all TG pairs as co-function ones.

Next, for a pair of TFs, we compute three TG coherence scores in terms of co-expression, co-function and interaction. The TG co-expression coherence score is computed as follows. We construct two sets of TGs: the first set includes all TGs that are regulated by both TFs; the second set is the reference set and includes all possible TGs that are regulated by any TF in the genome. For each TG set, we compute the fraction of TG pairs within the set that are co-expressed. The TG co-expression coherence score is then defined as the ratio of these two fractions. We then threshold this score with the cutoff >2. The TG coherence scores for other relationships (co-function and interaction) are computed in a similar way.

In total, we collected 9 TG coherence features measuring TG co-expression, co-function and interaction based on ChIP–chip (Features F7, F8 and F9 in Figure 2), literature (Features F10, F11 and F12 in Figure 2) and motif occurrence based (Features F13, F14 and F15 in Figure 2) transcriptional regulatory datasets. As shown in Figure 1, all features are strong predictors for TF cooperativity (likelihood ratio >4, Fisher's exact test P-value $< 10^{-5}$).

Bayesian network method

We use a Bayesian network framework to predict TF cooperativity by integrating TF pair features. For each TF pair, the prediction of cooperativity is based on the calculation of the posterior odds of cooperativity given the presence of genomic features. The posterior odds for predicting the class label y (1 if cooperativity exists, and 0 otherwise) by integrating genomic features $f_1, f_2, \ldots, f_n$ can be written as follows using the Bayes rule:

$$\log\frac{P(y=1|f_1,f_2,\ldots,f_n)}{P(y=0|f_1,f_2,\ldots,f_n)}=\log\frac{P(y=1)}{P(y=0)}+\log\frac{P(f_1,f_2,\ldots,f_n|y=1)}{P(f_1,f_2,\ldots,f_n|y=0)} \quad \mathbf{4}$$

where $y = 1$ represents TF cooperativity and $y = 0$ represents non-cooperativity. f_1 through f_n are different genomic features that are predictive for TF cooperativity. $P(y = 1 \mid f_1, f_2, \ldots, f_n)$ is the probability that the TF pair is cooperative given these features. $P(y = 1)/P(y = 0)$ is the prior odds. $P(f_1, f_2, \ldots, f_n | y = 1)/P(f_1, f_2, \ldots, f_n \mid y = 0)$ is the likelihood ratio for the combined features. A TF pair is predicted to be cooperative if the calculated posterior odds of cooperativity is greater than a predetermined threshold.

There are two special cases to the above general Bayesian network formalism. The first special case is naïve Bayes, where genomic features are assumed to be conditionally independent given TF cooperativity. In this case, the likelihood ratio of the combined features is equal to the product of the likelihood ratios for individual features. The second special case is full Bayesian Network, where none of the features are conditionally independent. In this case, the predictive power of all possible combinations of features values must be estimated. In general, the best Bayesian network structure lies somewhere between these two special cases. It is possible to learn the optimal Bayesian network structure from training data, but this problem is hard in terms of computational complexity, and requires a large training data (33,52,53). In this article, we rely on prior knowledge to determine the Bayesian network structure. The structure is then fixed during training and testing. This way, the computational complexity of the problem is dramatically reduced, and only a small training set is required.

In determining the Bayesian network structure, we used the guiding rule that the structure should be as simple as possible, i.e. maximize the number of conditional independencies among features, while at the same time still be able to capture the dominant dependencies within data. The final Bayesian network structure is shown in Figure 2. This Bayesian network is similar to naïve Bayes in that almost all features are conditionally independent. The only difference is that our Bayesian network takes into account the additional strong redundancy between TG-based features such as TG overlap and TG coherence. The rationale is explained in detail in the 'Results' section and the Supplementary Data.

Given the Bayesian network structure in Figure 2, we can determine the posterior odds in Equation (4) for every TF pair:

$$\log \frac{P(y = 1 | f_1, f_2, \ldots, f_n)}{P(y = 0 | f_1, f_2, \ldots, f_n)} = \log \frac{P(y = 1)}{P(y = 0)} + \sum_{i=1}^{n} \log \frac{P(f_i | y = 1, S_i)}{P(f_i | y = 0, S_i)} \quad (5)$$

Where S_i is the set of parent features that f_i conditionally depends upon.

RESULTS

Feature collection and assessment for predicting TF cooperativity

We collected 174 TFs (Supplementary Table 1) in yeast and 15 TF pair features that potentially correlate with TF cooperativity relationships. The 15 TF pair features include physical/genetic interaction, co-expression and co-evolution relationships among TFs, as well as the degree of overlap and coherence among the corresponding TGs in terms of co-expression, co-function and interaction, based on literature, ChIP–chip and motif occurrence evidence. Our feature collection is based on three insights: first, cooperative TFs tend to co-express, co-evolve and interact; second, cooperative TFs tend to share larger number of TGs than expected by chance; third, the TGs of cooperative TFs tend to share significant co-expression, co-function and interaction relationships (see Supplementary Figure 1 for a schematic outline; detailed description of data sources and characterization of all features can be found in 'Materials and Methods' section, Supplementary Data and Supplementary Figures 2–6).

We quantitatively assess the usefulness of each feature for TF cooperativity prediction by calculating the likelihood ratio scores using known transcriptional complexes as gold-standard. All 15 genomic features are good predictors for TF cooperativity (Figure 1). All features except TF co-expression have likelihood ratio scores significantly larger then 4 (Fisher's exact test P-value $<10^{-5}$). TF co-expression feature has a likelihood ratio of 1.8 which is still statistically significant (Fisher's exact test P-value <0.05) with good coverage. In particular, our newly introduced TG overlap features have excellent predictive accuracy and coverage for TF cooperativity (Figure 1 and 'Materials and Methods' section).

Integrated prediction of TF cooperativity by Bayesian network

All 15 genomic features collected above are good predictors for TF cooperativity (Figure 1). To further improve TF cooperativity prediction, we integrate these 15 features using a Bayesian network framework. There are two crucial steps for Bayesian network integration. The first step is to determine the structure of the network which encodes the conditional (in)dependence relationships among the features. The second step is parameter estimation by computing the contingency tables associating each feature with its immediate parent features. The first step is difficult, whereas the second step is straightforward.

We rely on prior knowledge to determine the Bayesian network structure according to two criteria: first, choose Bayesian network structures that capture the dominant dependencies within data; second, choose simple structures over complex structures with similar predictive power. The simplest Bayesian network is called naïve Bayes, where all features are assumed to be conditionally independent, i.e. different TF pair features are independent within cooperative TF pairs, as well as within non-cooperative TF pairs. This is the preferred network structure when the conditional dependencies among features are not strong.

In this article, we assume approximate conditional independence for most of our genomic features. First, we make the reasonable assumption that TF-based features are conditionally independent of TG-based features.

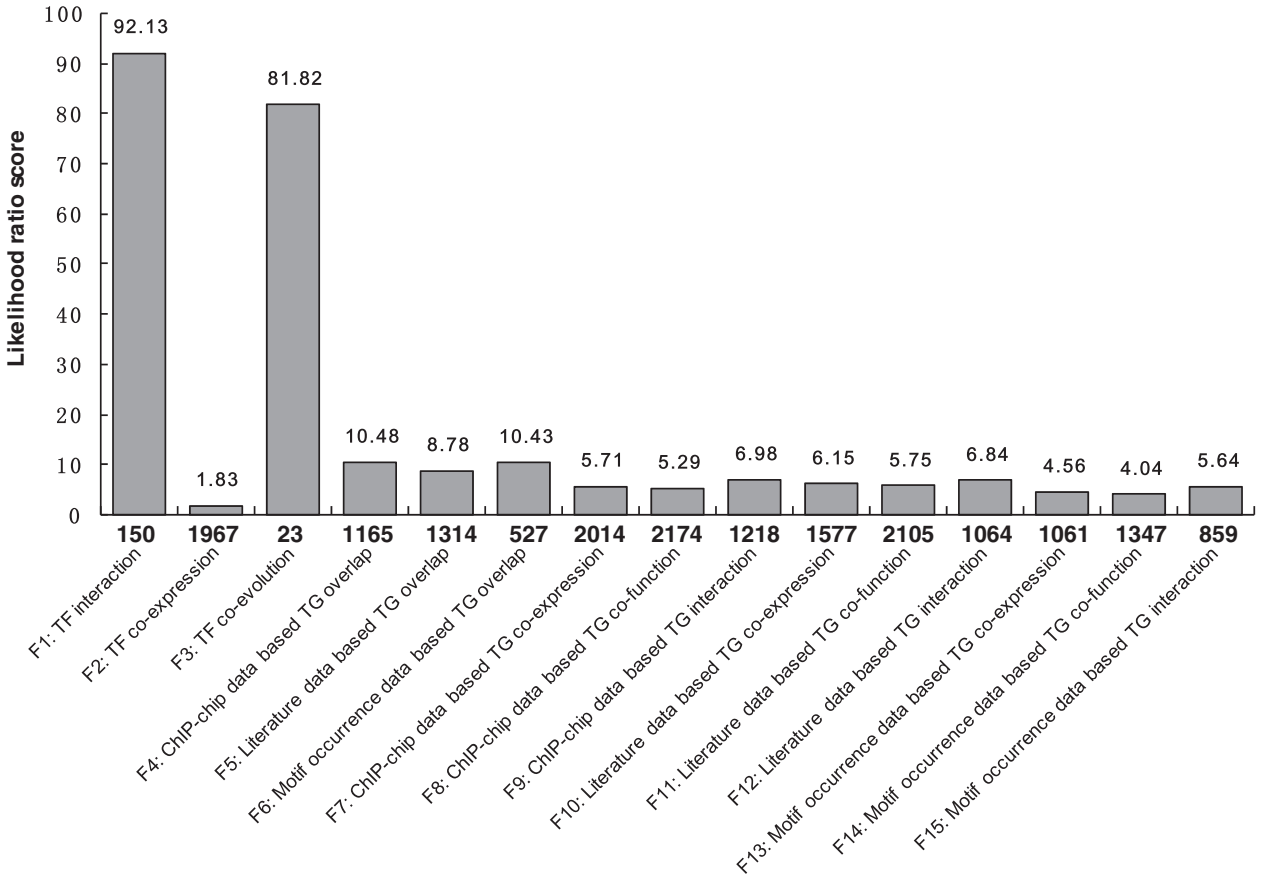

Figure 1. Likelihood ratio scores of the 15 genomic features for predicting transcriptional cooperativity. The *X*-axis represents the name of the features, and the *Y*-axis represents the likelihood ratio score associated with the presence of each feature. (The likelihood ratio scores associated with the absence of each feature is generally close to 1, and are not plotted here.) The features include TF physical/genetic interaction (F1), TF co-expression (F2), TF co-evolution (F3), TG overlap based on ChIP–chip (F4), literature (F5) and motif occurrence (F6), TG co-expression based on ChIP–chip (F7), literature (F10) and motif occurrence (F13), TG co-function based on ChIP–chip (F8), literature (F11) and motif occurrence (F14), TG interaction based on ChIP–chip (F9), literature (F12) and motif occurrence (F15). For each feature, the likelihood ratio score is above the bar and the number of predicted TF cooperativities is below the bar. A likelihood ratio score much larger than 1 indicates that the feature is a good predictor for TF cooperativity. The figure shows that TF interaction and co-evolution features are very strong predictors with limited coverage, TF co-expression feature is a weak predictor with broad coverage, and TG overlap and TG coherence features are strong predictors with good coverage.

Furthermore, the TF-based features such as TF interaction, co-expression and co-evolution, are based on measurements at the protein level, gene-expression level and sequence level respectively. They can be assumed to be approximately conditionally independent of each other. Second, TG-based features can be computed separately from large-scale ChIP–chip, small-scale experiments and motif occurrence based transcriptional regulatory datasets. Though all depicting TF–TG interactions, these datasets are assembled from different sources, and it is reasonable to assume that they are approximately conditionally independent for predicting TF cooperativity. Third, TG coherence can be measured in terms of co-expression, co-function and interaction. We make the reasonable assumption that these three features are conditionally independent. Overall, these data sources are not related to each other except for the fact that they are all good predictors for TF cooperativity. As a result, the conditional independence assumptions are approximately valid.

On the other hand, there is further redundancy among TG overlap features and TG coherence features that goes beyond the fact that they are all good predictors for TF cooperativity. In other words, TG overlap features and TG coherence features are not independent conditioning upon TF cooperativity. This is because TG overlap features and TG coherence features both rely on the TG information, and they measure different aspects of the set of co-regulated TGs: for a TF pair, TG overlap features measure the significance of the overlap between two TG subsets, while TG coherence features measure the level of coherence within the set of overlapping TGs in terms of the enrichment of co-expression, co-function and interaction relationships. Indeed, there is a large degree of redundancy. For example, in the literature based transcriptional regulatory dataset, TG overlap feature predicts 1314 cooperative TF pairs, among which 19 are in GSP. TG co-expression feature predicts 1577 cooperative TF pairs, among which 16 are in GSP. There is a significant overlap between these two sets: there are 535 shared

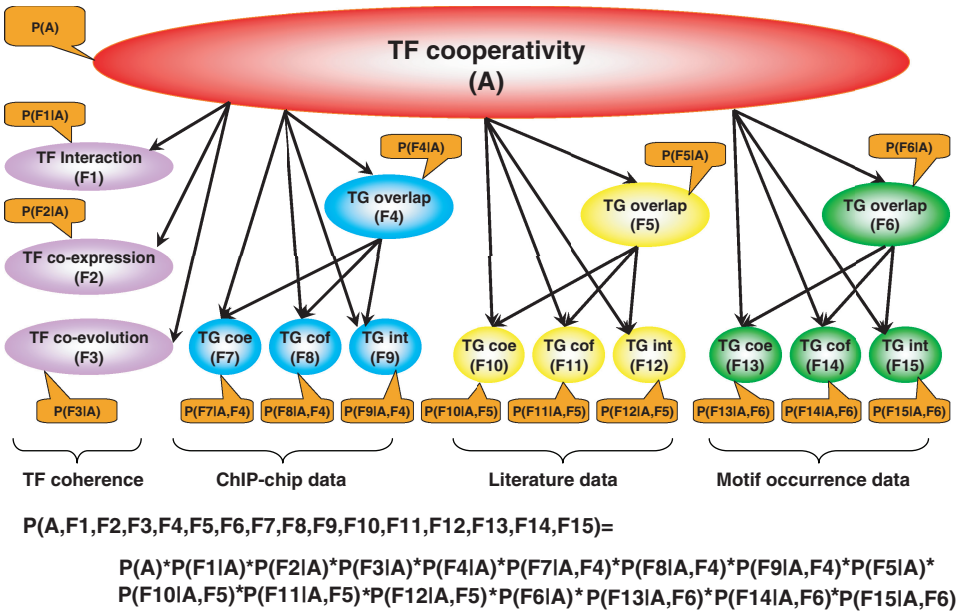

Figure 2. The architecture of the Bayesian network for TF cooperativity prediction. The following abbreviations are used (TF: transcription factor; TG: target gene; coe: co-expression; cof: co-function; int: interaction). Each node represents a particular genomic feature, and is labeled by their abbreviation name (F1–F15, see Figure 1 legend for details), as well as by the conditional probability for the random variable [for example P(F1|A)]. Each edge represents a direct conditional dependence between two variables. The equation for calculating the joint probability of 15 genomic features and TF cooperativity using this Bayesian network is specified at the bottom of the figure. From this joint probability one can then compute the conditional probability of TF cooperativity given genomic features. The rationale for choosing this Bayesian network architecture is given in the Supplementary Data.

predictions (Pearson's Chi-squared test P-value $<10^{-4}$), among which 16 are in GSP. As a result, the simplest naïve Bayes framework does not work well here.

On the basis of the above analysis, we constructed a simple yet reasonable Bayesian network architecture that takes into account the inter-dependencies of different TG-based features (Figure 2). In this network, features based on TG coherence are conditionally independent given TF cooperativity and TG overlap, whereas all other features are conditionally independent given TF cooperativity only. This architecture is the simplest that still captures the two types of dominant relationships among features and TF cooperativity: first, each feature alone is a good predictor for TF cooperativity; second, there is a strong correlation between TG overlap and TG coherence features, given TF cooperativity. We used contingency tables (Supplementary Tables 3–6) to compare our Bayesian network structure with alternative structures (see Supplementary Data for details). We found that our Bayesian network structure outperforms naïve Bayes because it takes into account the conditional dependencies among features that are important for predicting TF cooperativity. Second, our Bayesian network structure outperforms the full Bayesian network by dramatically reducing the number of parameters to be estimated and allowing accurate determination of all parameters without over-fitting the training data. Third, our Bayesian network model is relatively simple with a pre-determined structure, and we are only learning Bayesian network parameters but not structure from data. As a result, our small set of GSPs is sufficient for accurate learning.

Once the Bayesian network structure is fixed and the parameters are estimated, subsequent statistical inference is straightforward. To predict if a TF pair is cooperative or not, we first calculate conditional likelihood ratios for each feature given the parent features in the Bayesian network (Figure 2). Then we compute a total score by summing up the natural logarithm of all the conditional likelihood ratios associated with different features ('Materials and Methods' section). The TF pair is predicted to be cooperative if the total score exceeds a given threshold.

We estimate the true positive rate (TPR; fraction of GSP that are correctly predicted; sensitivity) and the false positive rate (FPR; fraction of GSN that are incorrectly predicted; 1-specificity) of different classifiers using 5-fold cross-validation, and plot the receiver operating characteristic (ROC) curves of sensitivity versus 1-specificity for our Bayesian network classifier, the naïve Bayes classifier based on each of the features, as well as the naïve Bayes classifier based on the integration of all features in Figure 3. In addition to 5-fold cross-validation, we also performed leave-one-out cross-validation and obtained similar true positive and FPR estimates.

Every feature is predictive (likelihood ratio >2 for all cases; see Figure 1), yet the ranking of predictive power for individual features is different at different FPR levels (Figure 3). For example, when the FPR is low (<0.01), the best individual features are TF interaction and TF co-evolution. When the FPR is high (>0.03), the best individual features are TG overlap ones. When the FPR is in the mid-range, the best individual features are TG

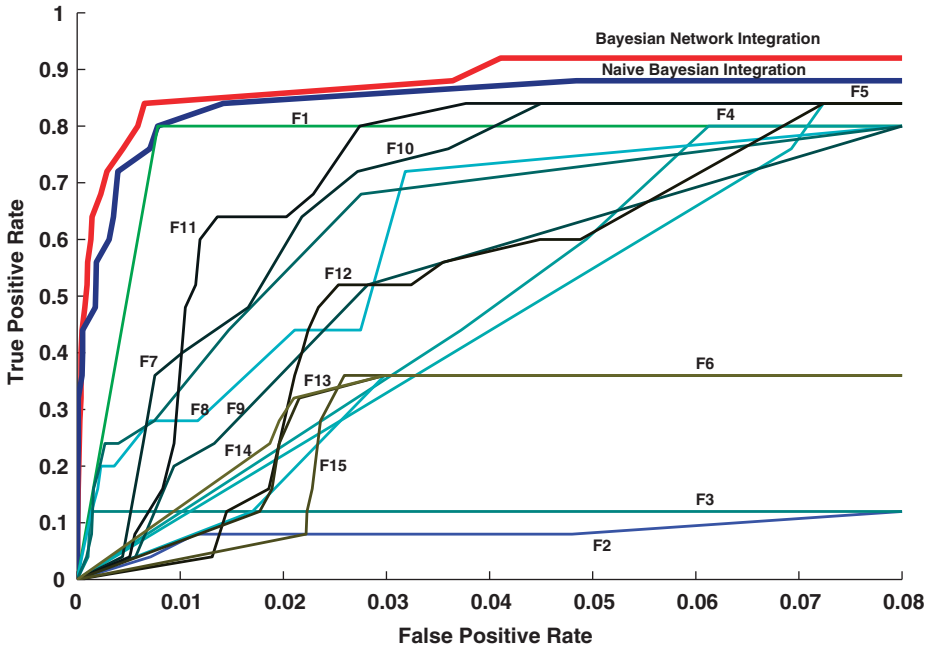

Figure 3. ROC plot comparison of our Bayesian network, naïve Bayes and 15 individual feature based classifiers. The 15 features (F1–F15) are explained in detail in Figure 1 legend.

coherence ones. Even though a reduced set of features will make it feasible to systematically search for the best network structure, and will make it easier to extend our current methodology to other organisms, in our case these features are complementary to each other, and the best predictions are generated by integrating all these features together, rather than selecting a minimal set of features. Indeed, our Bayesian network classifier significantly and consistently outperforms any of the individual classifiers at all FPR levels. Moreover, our Bayesian network classifier outperforms feature integration using naïve Bayes classifier (Figure 3), as well as logistic regression, decision tree, and *k*-nearest neighbor methods (these three methods are implemented by data mining software Orange (54); see Supplementary Figure 9). In particular, our Bayesian network method consistently outperforms naïve Bayes in terms of the area under the ROC curve (AUC) score, both in leave-one-out cross-validation, and in 98 out of 100 independent 5-fold cross-validation simulations (P-value < 0.05). The improved accuracy is due to three reasons. First, our Bayesian network is able to capture the important conditional dependencies among heterogeneous data in a graphical model (Figure 2). Second, parameter estimation in our Bayesian network requires much fewer training data than most other classification models. This is especially important in our case where the training set is very small. Third, Bayesian network is different from many machine learning methods in that feature dependencies are explicitly specified in a fully probabilistic way.

Reconstructed TF cooperativity network in yeast

To construct our final integrated classifier, we need to choose a cutoff: we predict a TF pair to be cooperative if the summed log-conditional likelihood ratio for TF cooperativity is greater than the cutoff, and non-cooperative if otherwise. This can be done in several ways (28,55). For example, this cutoff can be chosen to be the negative of the prior log-odds of observing cooperativity for a random TF pair. Unfortunately, it is difficult to estimate this prior log-odds, because we do not know the exact number of cooperative TF pairs. Alternatively, in this article our final predictions for TF cooperativity is based on thresholding the positive predictive value (PPV), defined as the fraction of positive predictions that are in GSP. In Supplementary Figure 10, we plot the sensitivity (TPR), specificity (1 – FPR), PPV and the percentage of positive predictions under different cutoff choices. We chose 7.34% as our final cutoff of PPV, i.e. 7.34% of the predicted cooperative TF pairs belong to the GSP set. At this PPV cutoff, our final integrated prediction consists of 286 cooperative relationships among 113 TFs (sensitivity 0.84, specificity 0.982). In Supplementary Table 7, we list all 286 predicted interactions together with their Bayesian network integration scores. Among these, 21 are in the GSP set. The fraction of GSP set correctly predicted by our Bayesian network classifier is 84%. Training our Bayesian network classifier on the entire gold-standard dataset yields true positive and FPRs (TPR = 84%, FPR = 1.8%) that are similar to 5-fold and leave-one-out cross-validation based estimates (TPR = 84%, FPR = 1.4% and 1.5%, respectively), suggesting that model overfitting is not an issue here. This is further supported by additional assessment with two independent benchmark datasets (see section 'Comparison with other methods' below).

To get a high-confidence subset of TF cooperativity predictions, we choose an even more stringent PPV

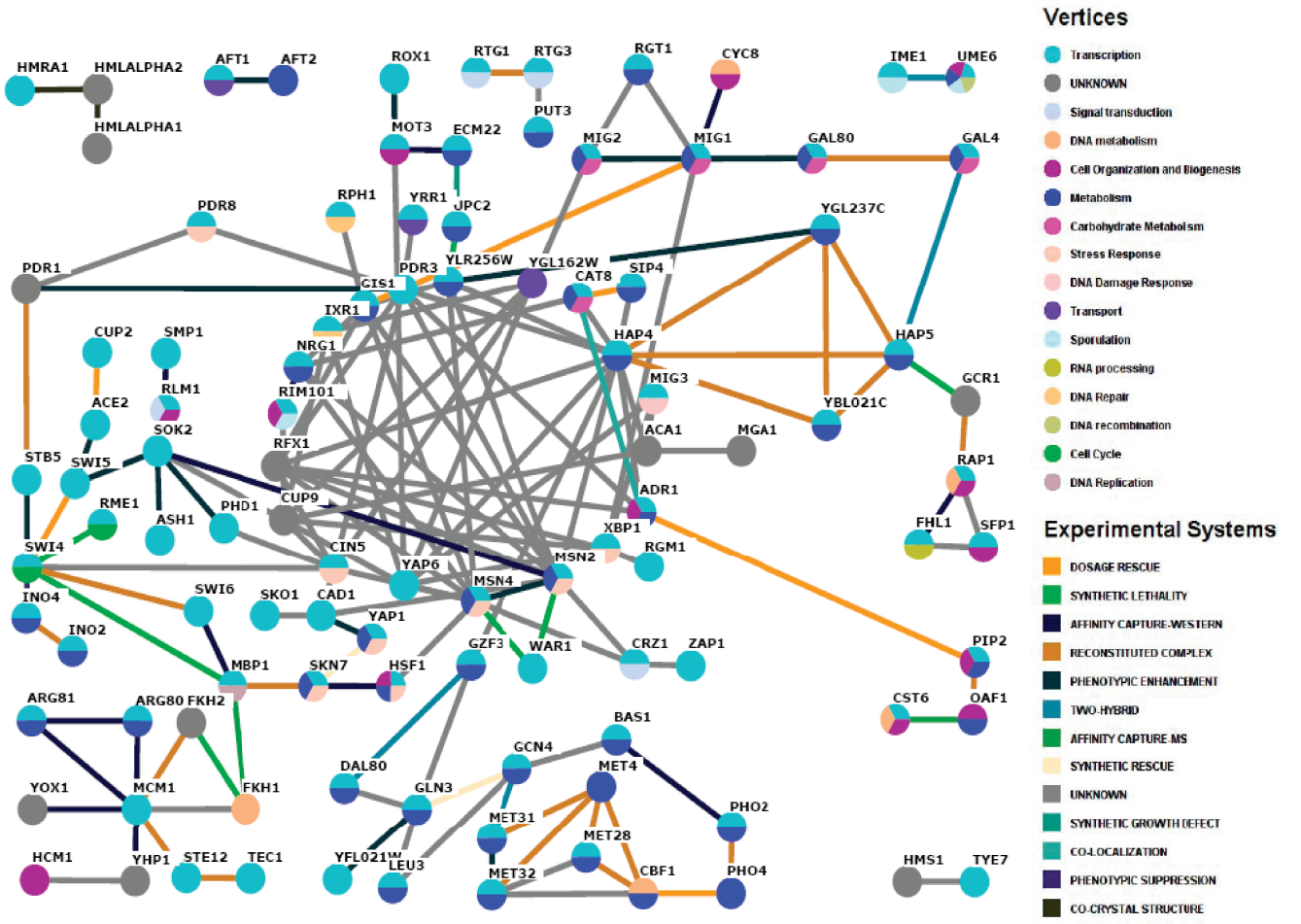

Figure 4. A map of the predicted high-confidence TF cooperativity network in yeast. Nodes represent TFs and edges represent cooperativity relationships among them. The node colors represent GO function annotation (grey represents unknown function). The edges are the top 159 TF cooperativity predictions among 105 TFs. TF–TF interactions that also appear in BioGRID are colored by their experimental method of detection. The newly predicted TF cooperative relationships not in BioGRID are colored by grey. The figure is generated by Osprey (75).

cutoff of 13.21% (sensitivity 0.84, specificity 0.992). Using this stringent cutoff, we obtain 159 predicted cooperative relationships among 105 TFs. In Figure 4 we visualize this high-confidence yeast TF cooperativity subnetwork. We further label TFs with their GO functional annotations, and label TF–TF interactions that appear in BioGRID with the experimental method of detection. Importantly, we are able to predict many new TF cooperative relationships that are not in the BioGRID (Figure 4). Overall, our novel predictions (other than those that overlap with GSPs) can be grouped into two categories: (i) new cooperativity relationships in and around known transcriptional complexes and (ii) other novel cooperative relationships.

To further validate our prediction, we searched the PubMed database and manually curated TF cooperativity information from literature abstracts. We found that most of the 159 predicted TF cooperativity relationships are supported by one or more published literatures (143 out of 159 are supported by literature evidence including 21 GSPs). We made sure that these new literature-based evidences were not included in the feature collections for prediction, thus the validation step is completely independent of the training step. The extensive literature validation demonstrates the overall high quality of the prediction results. In Supplementary Table 8, we list all 159 high-confidence TF cooperativity predictions with their Bayesian network integration score, together with detailed descriptions of literature and experimental evidences for TF–TF physical, genetic and regulatory interaction, as well as other documented TF cooperativity evidences.

Literature validation of TF cooperativity predictions

We predict the cooperativity between MATa1 and MATα2 based on TF interaction and TG overlap evidences (Supplementary Data). This TF pair plays an important role in determining yeast cell type by forming a heterodimer that binds DNA and represses transcription in a cell-type specific manner (25,26). Whereas the α2 and a1 proteins on their own have only modest affinity for DNA, the α2/a1 heterodimer binds DNA with high

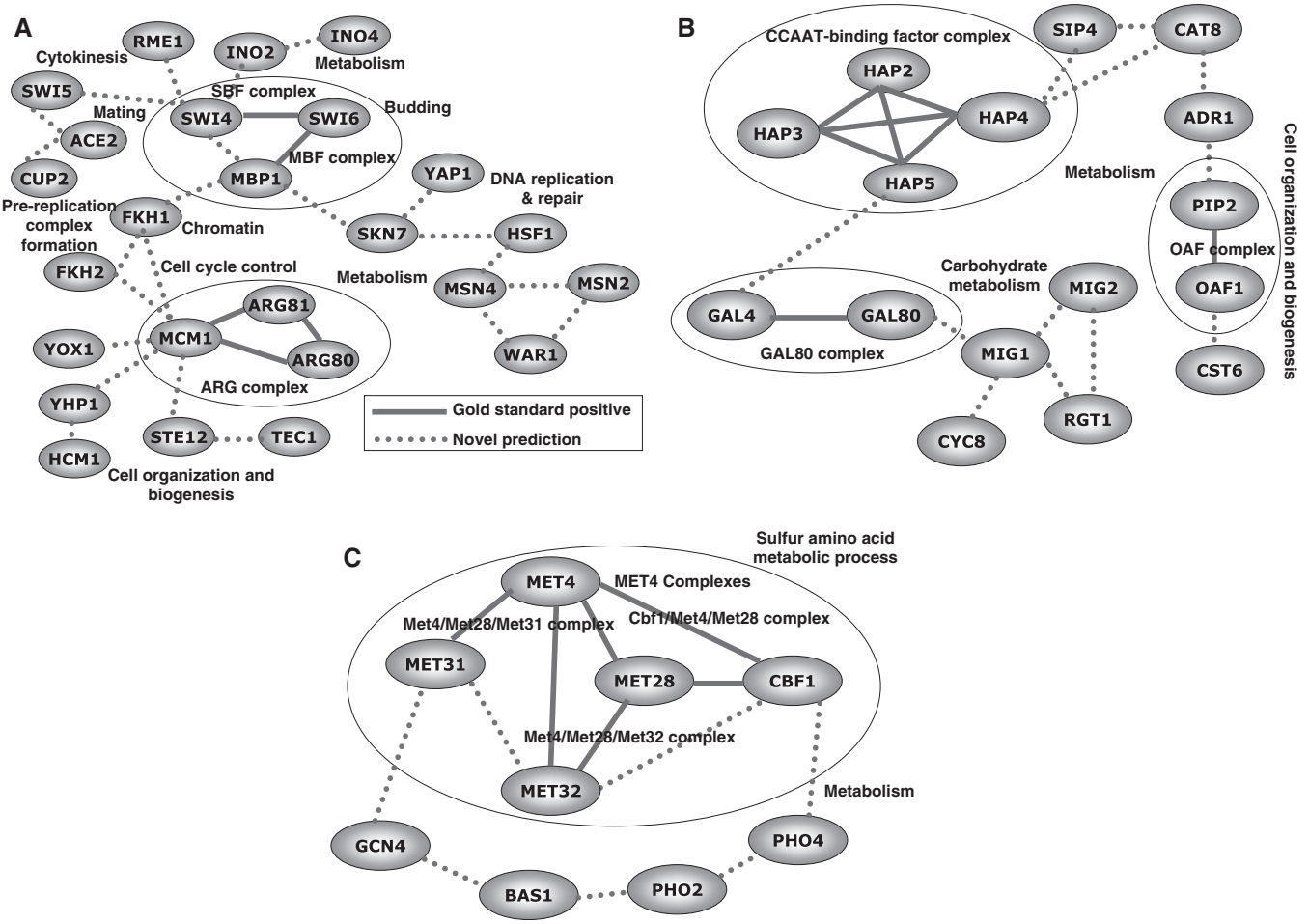

Figure 5. Three predicted transcriptional complexes by our method. (**A**) Transcriptional complex related to cell-cycle process and arginine metabolic process. (**B**) Transcriptional complex related to carbohydrate metabolic process and galactose metabolic process. (**C**) Transcriptional complex participating in the regulation of sulfur metabolism. Known co-complexation relationships in the GSP dataset are represented as solid lines. Novel predictions are represented by dashed lines and are validated by literature in the Supplementary Data and Supplementary Table 8.

specificity and affinity, preferring its own binding site over random DNA by a ratio of at least 10^5. The three dimensional crystal structure of the α2/a1 heterodimer bound to DNA was determined at a resolution of 2.5 Å (Supplementary Figure 11).

We predict three large transcriptional complexes in yeast (Figure 5). Figure 5A illustrates TF cooperativity predictions that are related to cell-cycle control. We predict the cooperativity between Mbp1 and Swi4. Mbp1 and Swi4 share 50% sequence identity in their DNA-binding domains. For many G1/S-regulated genes, removal of both Swi4 and Mbp1 was necessary and sufficient to essentially eliminate cell-cycle-regulated expression (56). Although some level of redundancy exists between Mbp1 and Swi4, there is extensive experimental support that Mbp1 and Swi4 regulate different subsets of genes involved in distinct biological processes (37,57), and that the SBF complex (formed between Swi4 and Swi6) act in concert with the MBF complex (formed between Mbp1 and Swi6) in regulating late G1-specific transcription (58). We predict that ARG transcriptional complex (Arg80, Arg81 and Mcm1) is functionally related to SBF and MBF complexes through Fkh1 and Fkh2. The two forkhead TFs, Fkh1 and Fkh2, are known to cooperate with Mcm1 to control M-phase transcription (59). Mcm1 facilitates Fkh1 binding to DNA, and they jointly regulate recombination enhancer activity in **a**-cell (60). In addition, the DNA-binding domains of Fkh1 and Fkh2 share 72% sequence identity, and the double mutant of Fkh1 and Fkh2 displays obvious morphological change (59). Fkh1-Mbp1 cooperativity is detected by affinity capture-MS experiment as physical interaction (61). In addition to the two complexes and the cooperative connections between them, we also make many novel TF cooperativity predictions in the neighborhood of these two complexes (Figure 5A), all of which are validated by literature (Supplementary Data).

Figure 5B shows the TF cooperativity predictions that are related to oxidative metabolism and carbohydrate metabolism. The predicted cooperativity between Hap5 and Gal4 links the known CCAAT-binding factor complex (Hap2/Hap3/Hap4/Hap5) and GAL80 complex (Gal80/Gal4), and is supported by yeast two-hybrid experiment (62) and the fact that their binding motifs

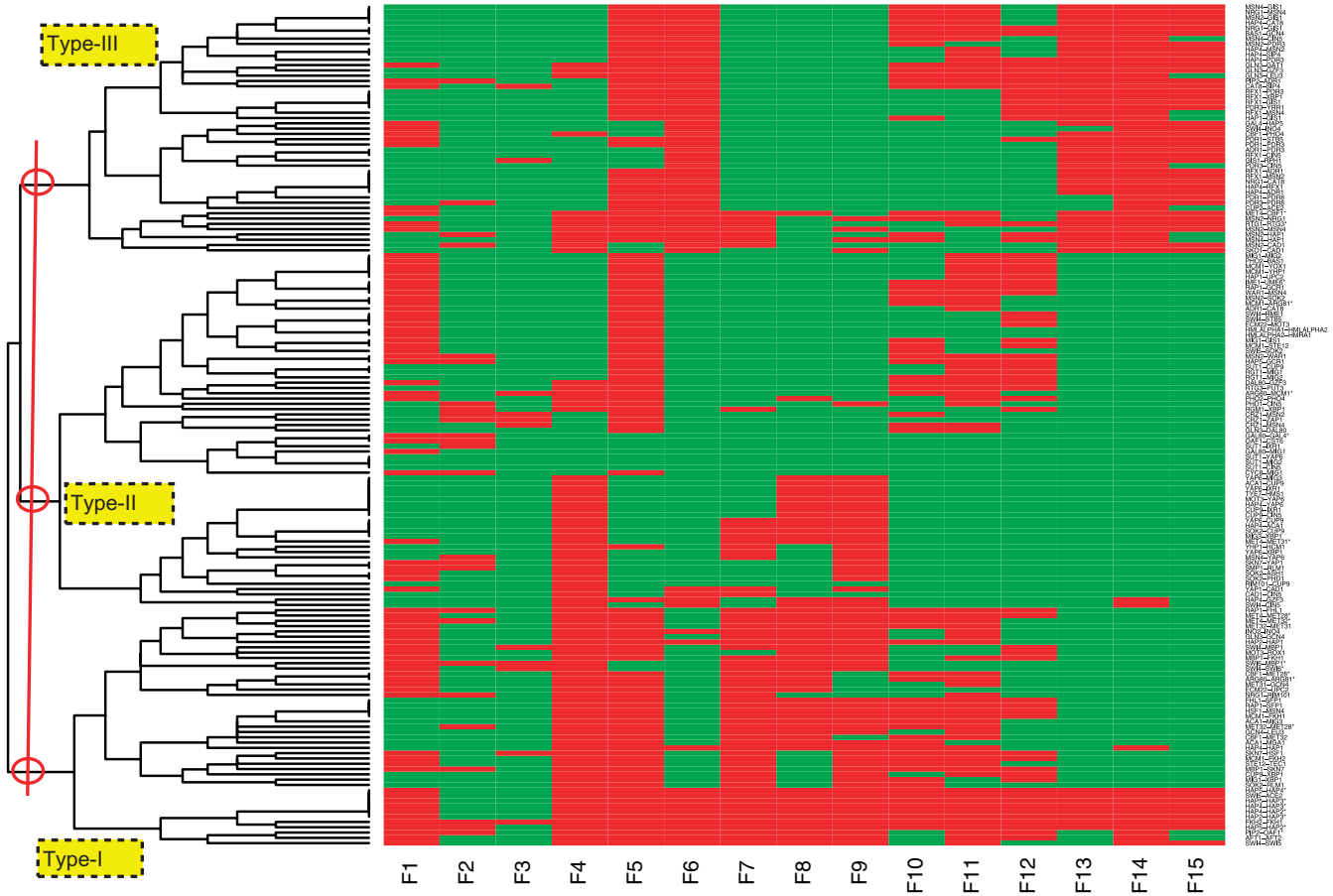

Figure 6. The heatmap for feature profiles of the predicted 159 TF cooperative relationships. Columns represent genomic features (the meaning of the feature names are given in Figure 3 legend), and rows represent predicted cooperative TF pairs. Cells are colored by red if the feature is present ($f = 1$), or green if the feature is absent ($f = 0$). The cooperative TF pairs in the GSP dataset are highlighted by asterisk symbol. The predicted TF cooperativities are grouped into three types following hierarchical clustering of the feature profile: type-I (45 TF pairs), type-II (67 TF pairs) and type-III (47 TF pairs).

co-occur in the promoter regions of functionally related genes Fkh1, Fkh2 and Mot3. The predicted cooperativity between Adr1 and Cat8 bridges the OAF complex (Pip2/Oaf1) and CCAAT-binding factor complex (Hap2/Hap3/Hap4/Hap5), and is supported by the experimental evidence that they synergistically activate the glucose-regulated alcohol dehydrogenase gene Adh2 (63) (Supplementary Data).

In Figure 5C, the predicted TF cooperative pairs are located near the Met4 complex which regulates sulfur metabolism and oxidative stress response genes. We predict two more novel TF cooperativity relationships (Met31-Met32 and Met32-Cbf1) inside the Met4 complex. Furthermore, we identified several additional TFs that are cooperatively linked to the Met4 complex: Gcn4, Bas1, Pho2 and Pho4. In particular, the predicted cooperativity between Pho4 and Cbf1 is supported by interaction, motif occurrence, and literature evidences (64). All of these new predictions are supported by varying degrees of literature evidence (Supplementary Data).

Three types of TF cooperativity

We performed a global analysis on the different feature combinations for the top-scoring cooperative TF pairs (see Supplementary Data). The dendrogram of hierarchical clustering on the feature profile (Figure 6) reveals three different types of predicted TF cooperativities. In type-I cooperativity (45 TF pairs), the TF–TF physical interaction feature is the dominant predictor, and most other features are useful predictors as well. Most GSP data belong to type-I cooperativity (14 out of 21). In type-II cooperativity (67 TF pairs), TF cooperativity is supported largely by TF–TF genetic interaction and ChIP–chip/literature based transcriptional regulatory information, but not by motif occurrence information, primarily due to distinct DNA-binding motifs for the pair of TFs. In type-III cooperativity (47 TF pairs), TF cooperativity is supported largely by motif occurrence transcriptional regulatory information, primarily due to similar DNA-binding motifs for the pair of TFs, and in general not supported by TF–TF interaction information. The three different patterns of feature combination suggest three possible types of TF cooperativity mechanisms in transcriptional control (Supplementary Figure 12).

In type-I cooperativity, TFs cooperate through physical interactions in a transcriptional complex, and jointly regulate many TGs. Upon the formation of the complex, the binding probability of RNA polymerase onto the

promoter sequence will either increase or decrease, thus affecting the subsequent transcriptional efficiency. Here protein physical interaction is the dominant predictor for this kind of cooperative relationships. Examples of this type of cooperativity include the Met4 Complex, CC AAT-binding factor complex, MBF complex and SBF complex. Since the two TFs cooperate in a transcriptional complex and they stay together most of the time, we expect that other features are good predictors as well. For example, the co-evolution relationship is a good predictor since interacting TFs tend to co-evolve in different genomes.

In type-II cooperativity, TFs cooperate largely through genetic or regulatory interaction, and jointly regulate many TGs. Different from the first case, TFs interact in a genetic or regulatory module that may or may not be explained by the existence of a physical complex. For example, one TF can regulate a secondary TF, and they jointly regulate the TGs. The well-known network motifs such as the feed-forward loop and the regulator cascade (37) belong to this type. Alternatively, one TF can genetically interact with another TF which leads to a joint phenotype. Here, TF regulatory or genetic interactions are the dominant predictors for type-II cooperativity. Examples of this type of cooperativity include Sok2-Phd1, Yap6-Cup9 and Skn7-Yap1 pairs, which are all detected by ChIP–chip experiments as belonging to regulatory feed-forward motifs (37) and supported by literature evidences (see Supplementary Data).

In type-III cooperativity, TFs often cooperate with each other in a competitive way (TF competitive regulation), or in a redundant manner. In this case, the DNA-binding sites of these TFs share high sequence identity or similarity. Their binding motifs often overlap with each other, leading to competition between TFs for transcriptional control. Here, the motif occurrence data based features are the dominant predictors (Figure 6). Some representative cases of type-III cooperativity such as Pdr1-Pdr3, Cat8-Sip4 and Msn2-Msn4 are well supported by literature (Supplementary Data). Competitive regulation is an important kind of functional interaction among TFs; the interplay between cooperation and competition is a powerful mechanism to achieve complex regulatory control (65).

Comparison with other methods

In what follows, we extensively compare our TF cooperativity predictions with the following five existing methods. (i) Jansen *et al.* (28) used Bayesian networks to predict yeast protein–protein interactions in general, by integrating genomic features such as mRNA co-expression, coessentiality, co-localization and experimental interaction datasets. The predictions of Jansen *et al.* (PIT, with likelihood ratio cutoff 300) contain 37 putative TF–TF interactions. (ii) Datta *et al.* (19) used log-linear models to predict cooperative binding among cell cycle specific TFs. Their top 25 cooperative TF pairs regulating cell-cycle processes are used for comparison. (iii) Banerjee *et al.* (20) integrated genome-wide location data from ChIP–chip and gene-expression data to infer 183 cooperative TF pairs by expression correlation of the TGs. (iv) Tsai *et al.* (22) used statistical methods (ANOVA) to identify synergistic pairs of yeast cell-cycle TFs by combining ChIP–chip data and microarray data, and generated three sets of predictions (confident, doubtful and plausible). (v) Balaji *et al.* (9) used a specific network transformation procedure to obtain a co-regulatory network describing the set of all significant associations among TFs in terms of regulating common TGs, and generated two sets of predictions (core and all).

We assess the quality of these TF cooperativity predictions by calculating their overlap with two independent, high-quality benchmark datasets that are not used in these methods. The first benchmark dataset is based on the KEGG pathway database (66), and contains 48 TF pairs among 13 TFs that co-occur in at least one KEGG pathway. The second benchmark dataset is based on the recently published high-quality experimental binary protein–protein interaction map in yeast (CCSB-YI1) by Yu *et al.* (67), and contains 17 interacting TF pairs among 24 TFs. In Table 1, we compare the overlap of different predictions with these two benchmark datasets. Our predictions overlap with both KEGG and CCSB-YI1 datasets more significantly than all other existing predictions (Fisher's exact test *P*-values in Table 1). This shows that our TF cooperativity predictions are better in quality than existing methods.

Our predictions compare favorably to existing methods due to at least one of the following reasons: (i) many features useful for predicting protein–protein interactions in general in (28), such as co-expression, coessentiality and co-localization, are not as useful for predicting TF cooperativity; (ii) we compiled new features that are specifically useful for predicting TF cooperativity, such as TG overlap and TG coherence; and (iii) we integrated many different features to predict TF cooperativity on a genomic scale.

At the same time, we also found significant overlap between our predictions and results based on all existing methodologies: Jansen *et al.* (six overlapping TF pairs, Fisher's exact test *P*-value $<10^{-4}$), Datta *et al.* (five overlapping TF pairs, Fisher's exact test *P*-value $<10^{-2}$), Banerjee *et al.* (20 overlapping TF pairs, Fisher's exact test *P*-value $<10^{-10}$), Tsai *et al.* (five overlapping TF pairs for 'doubtful' predictions, Fisher's exact test *P*-value $<10^{-2}$) and Balaji *et al.* (118 overlapping TF pairs for 'all' predictions, Fisher's exact test *P*-value $<10^{-13}$; 107 overlapping TF pairs for 'core' predictions with co-regulation coefficient greater than 1, Fisher's exact test *P*-value $<10^{-26}$). The overlaps, while statistically significant, are not large, most likely due to limited coverage of case-study predictions and predictions based on single data source, for a given quality cutoff. One main advantage of our method is the ability to increase prediction coverage by integrating diverse genome-wide data sources.

Overall, these comparisons further demonstrate the feasibility and effectiveness of our supervised learning approach applied here for the first time to the genome-wide, integrated prediction of the TF cooperativity network.

Table 1. Comparison of our Bayesian network method with existing methods

Benchmark Dataset	Our method	Datta *et al.* (19)	Banerjee *et al.* (20)	Tsai *et al.* (22) (doubtful)	Tsai *et al.* (22) (plausible)	Tsai *et al.* (22) (confident)	Balaji *et al.* (9) (all)	Balaji *et al.* (9) (core)	Jansen *et al.* (26)
KEGG pathway database (63) (13 TFs, 48 TF pairs)									
Number of overlapping TFs	13	4	8	7	8	6	13	13	7
Number of possible interactions among overlapping TFs	78	6	28	21	28	15	78	78	21
Number of KEGG interactions among overlapping TFs	48	6	20	15	18	9	48	48	13
Number of predicted interactions among overlapping TFs	8	3	8	3	1	2	69	48	2
Number of KEGG interactions that are correctly predicted	8	3	5	3	1	1	45	33	2
Fisher's exact test *P*-value	**0.016**	**1.0**	**0.87**	**0.34**	**0.64**	**0.86**	**0.071**	**0.079**	**0.37**
CCSB-YI1 dataset (64) (24 TFs, 17 TF pairs)									
Number of overlapping TFs	20	2	11	2	3	1	18	18	8
Number of possible interactions among overlapping TFs	190	1	55	1	3	0	153	153	28
Number of CCSB-YI1 interactions among overlapping TFs	13	0	2	0	0	0	12	12	3
Number of predicted interactions among overlapping TFs	5	1	2	0	1	0	91	50	2
Number of CCSB-YI1 interactions that are correctly predicted	5	0	0	0	0	0	3	3	1
Fisher's exact test *P*-value	**6.6×10^{-7}**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**0.82**	**0.21**

We compare our predictive results by Bayesian network method with eight predictive results by existing methods using two independent benchmark datasets: KEGG pathway dataset and CCSB-YI1 dataset. The statistical significance of the overlap between the prediction and the benchmark datasets is measured by the Fisher's exact test *P*-value.

DISCUSSION

Transcriptional cooperativity is a biologically rich and complex phenomenon that cannot be reduced to a simple, precise mechanistic definition. The same is true for other complex yet important biological relationships, such as genetic interaction and functional linkage. Here, we define TF cooperativity as the existence of any kind of functional interaction among TFs in regulating TGs. This working definition allows us to unify available notions of TF cooperativity. Under this broad definition, predicting transcriptional cooperativity becomes predicting functional associations between TFs, which is a sub-problem of the well-posed problem of predicting functional associations between genes. A potential drawback of this broad definition is the lack of high-resolution, mechanistic insights into how TFs cooperate in transcriptional regulation. Nevertheless, our comprehensive predictions of transcriptional cooperativity are useful for guiding the design of further experiments.

We introduced three machine-learning ideas for the first time into the prediction of transcriptional cooperativity. First, we introduced a small set of well-constructed gold-standard dataset, and emphasized its central role in our data integration framework. Second, we used graphical models such as Bayesian networks to capture the conditional dependence relationships among genomic features. The explicit specification of feature dependencies in a fully probabilistic way is especially important for our case, where the gold-standard data is scarce. Third, our Bayesian network structure is pre-chosen by considering the trade-off between predictive bias and variance, i.e. choosing the simplest structure possible, and only adding structural complexity when compelling biological justification exists. In this way, we only need to learn Bayesian network parameters during training (52,53). In this way, all parameters in the Bayesian network can be reasonably estimated using the small set of GSP data. In general, our methodology can be applied to other genomic data integration tasks where high-quality GSP data are scarce.

In our work, the gold-standard dataset plays a crucial role in assessing the predictive ability of each piece of evidence, in validating the superiority of our intuitively constructed Bayesian network structure compared to full Bayes and naïve Bayes, and in data integration for high-confidence prediction of TF cooperativity. Our GSP dataset is composed of mostly type-I cooperative TF pairs, as this is the only high-quality dataset that is currently available. This may result in underrepresentation of other types of transcriptional cooperativity in our predictions. On the other hand, our feature collections are systematic and extensive, and they are useful for predicting TF cooperativity in general. Indeed, the optimal classifier trained on the small, limited gold-standard dataset is able to generate biologically meaningful and literature-validated predictions for all three types of TF cooperativity.

Many cutoff choices in this article are heuristic. Such heuristics are necessary due to the lack of experimental data on transcriptional cooperativity. The majority of these cutoff choices follow standard practices in the field, and they make intuitive sense. The rest of the cutoffs are

chosen to ensure reasonable feature coverage. Our heuristic cutoff choices are not based on the class information of the gold-standard data, and the overall prediction results are robust to changes in these cutoff choices. To test the effect of cutoff choices on our predictions, we integrated the five most predictive features with the least heuristic cutoff choices, namely TF–TF interaction, co-evolution and the three TG overlap features, and left out all other features where more heuristic cutoff choices are used (TF co-expression and TG coherence features). We found that the predictive performance of our method drops, but not by too much: at the FPR level of 0.01, the TPR of the predictions based on the reduced feature set (0.76) is lower than that based on the full feature set (0.84), but still much higher than random predictions (0.01).

We rigorously validated our TF cooperativity predictions using independent data sources. We found significant overlaps between our predictions and previous computational predictions for TF cooperativity, as well as pathway databases such as KEGG. In addition, we validated most of our predictions using new PubMed literatures not used in our training procedure. These comparisons demonstrate the validity of our approach and the quality of our predictions.

The GSP set used in this article can be expanded in the future to include other well-characterized instances of TF–TF cooperativity, especially TF–TF genetic interactions and TF pairs co-regulating TGs. Larger training set will not only allow us to learn Bayesian network structure and parameters simultaneously, but also allow us to train different Bayesian networks that best predict and characterize different types of TF cooperativity.

Our results can be used to improve the accuracy of reconstructed transcriptional regulatory networks (68–70). In addition, our method can be extended to the prediction of cooperativity among three or more TFs, by looking for cliques in the reconstructed pairwise TF cooperativity map. Future directions also include extending our method to higher eukaryotes where TF cooperativity is expected to be more complex (13,16, 71,72), and to relate the alterations in these synergies to complex human diseases. Finally, given the conceptual similarity between TF- and microRNA-mediated control of gene expression, our method can be applied to study microRNA cooperativity (73,74), and more generally the cooperativity networks of any regulatory system in an organism.

CONCLUSIONS

In this article we reconstruct and analyze the cooperativity among TFs in yeast using Bayesian network integration of 15 diverse genome-wide data sources (sequence, expression, function, interaction and evolution) for both TFs and their corresponding TGs. To our best knowledge, this is the first time that the supervised learning framework is used to predict TF cooperativity. By assessing the predictive power of the individual features and integrating these diverse features within a Bayesian network framework, we constructed a high-confidence whole-genome map of predicted TF cooperativity in yeast. In addition, the existence and strength of the correlation between TF cooperativity and individual features, and the patterns of feature combination provide further biological insights into the different types of TF cooperativity.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank four anonymous reviewers for their helpful input.

FUNDING

National Natural Science Foundation of China (Grant No. 10801131 to Y.W.); Chinese Academy of Sciences (No. kjcs-yw-s7 to Y.W. and X.S.Z.); National Basic Research Program (No. 2006CB503900 to Y.W. and X.S.Z.); National Natural Science Foundation of China (Grant No. 10631070 and No. 60873205 to X.S.Z.); Research Starter Grant in Informatics from the PhRMA Foundation to Y.X. Funding for open access charge: National Basic Research Program (No. 2006CB503900).

Conflict of interest statement. None declared.

REFERENCES

1. Latchman,D.S. (1990) Eukaryotic transcription factors. *Biochemical J.*, **270**, 281–289.
2. Latchman,D.S. (1997) Transcription factors: an overview. *Int. J. Biochem. Cell Biol.*, **29**, 1305–1312.
3. Beer,M.A. and Tavazoie,S. (2004) Predicting gene expression from sequence. *Cell*, **117**, 185–198.
4. Buchler,N.E. (2003) On schemes of combinatorial transcription logic. *Proc. Natl Acad. Sci. USA*, **100**, 5136–5141.
5. Tan,K., Shlomi,T., Feizi,H., Ideker,T. and Sharan,R. (2007) Transcriptional regulation of protein complexes within and across species. *Proc. Natl Acad. Sci. USA*, **104**, 1283–1288.
6. Das,D., Banerjee,N. and Zhang,M.Q. (2004) Interacting models of cooperative gene regulation. *Proc. Natl Acad. Sci. USA*, **101**, 16234–16239.
7. Aguilar,D. and Oliva,B. (2008) Topological comparison of methods for predicting transcriptional cooperativity in yeast. *BMC Genomics*, **9**, 137.
8. Yu,X., Lin,J., Masuda,T., Esumi,N., Zack,D.J. and Qian,J. (2006) Genome-wide prediction and characterization of interactions between transcription factors in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, 917–927.
9. Balaji,S., Babu,M.M., Iyer,L.M., Luscombe,N.M. and Aravind,L. (2006) Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. *J. Mol. Biol.*, **360**, 213–227.
10. Bluthgen,N., Kielbasa,S.M. and Herzel,H. (2005) Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res.*, **33**, 272–279.
11. Tsong,A.E., Miller,M.G., Raisner,R.M. and Johnson,A.D. (2003) Evolution of a combinatorial transcriptional circuit a case study in yeasts. *Cell*, **115**, 389–399.

12. Tsong,A.E., Tuch,B.B., Li,H. and Johnson,A.D. (2006) Evolution of alternative transcriptional circuits with identical logic. *Nature*, **443**, 415–420.
13. Parisi,F., Wirapati,P. and Naef,F. (2007) Identifying synergistic regulation involving c-Myc and sp1 in human tissues. *Nucleic Acids Res.*, **35**, 1098–1107.
14. Wang,W., Cherry,J.M., Nochomovitz,Y., Jolly,E., Botstein,D. and Li,H. (2005) Inference of combinatorial regulation in yeast transcriptional networks: a case study of sporulation. *Proc. Natl Acad. Sci. USA*, **102**, 1998–2003.
15. Wagner,A. (1999) Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics*, **15**, 776–784.
16. Hannenhalli,S. and Levy,S. (2002) Predicting transcription factor synergism. *Nucleic Acids Res.*, **30**, 4278–4284.
17. Pilpel,Y., Sudarsanam,P. and Church,G.M. (2001) Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.*, **29**, 153–159.
18. Chang,Y.-H., Wang,Y.-C. and Chen,B.-S. (2006) Identification of transcription factor cooperativity via stochastic system model. *Bioinformatics*, **22**, 2276–2282.
19. Datta,D. and Zhao,H. (2008) Statistical methods to infer cooperative binding among transcription factors in Saccharomyces cerevisiae. *Bioinformatics*, **24**, 545–552.
20. Banerjee,N. and Zhang,M.Q. (2003) Identifying cooperativity among transcription factors controlling the cell cycle in yeast. *Nucleic Acids Res.*, **31**, 7024–7031.
21. Nagamine,N., Kawada,Y. and Sakakibara,Y. (2005) Identifying cooperative transcriptional regulations using protein-protein interactions. *Nucleic Acids Res.*, **33**, 4828–4837.
22. Tsai,H.-K., Lu,H.H.-S. and Li,W.-H. (2005) Statistical methods for identifying yeast cell cycle transcription factors. *Proc. Natl Acad. Sci. USA*, **102**, 13532–13537.
23. Yang,Y.-L., Suen,J., Brynildsen,M., Galbraith,S. and Liao,J. (2005) Inferring yeast cell cycle regulators and interactions using transcription factor activities. *BMC Genomics*, **6**, 90.
24. Wang,J. (2007) A new framework for identifying combinatorial regulation of transcription factors: a case study of the yeast cell cycle. *J. Biomedical Informatics*, **40**, 707–725.
25. Li,T., Jin,Y., Vershon,A.K. and Wolberger,C. (1998) Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucleic Acids Res.*, **26**, 5707–5718.
26. Li,T., Stark,M.R., Johnson,A.D. and Wolberger,C. (1995) Crystal structure of the MATa1/MATalpha2 homeodomain heterodimer bound to DNA. *Science*, **270**, 262–269.
27. Delcher,A.L., Kasif,S., Goldberg,H.R. and Hsu,W.H. (1993) Protein secondary structure modelling with probabilistic networks. *Proc. Int. Conf. Intelligent Sys. Mol. Biol.*, 109–117.
28. Jansen,R., Yu,H., Greenbaum,D., Kluger,Y., Krogan,N.J., Chung,S., Emili,A., Snyder,M., Greenblatt,J.F. and Gerstein,M. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**, 449–453.
29. Lee,I., Date,S.V., Adai,A.T. and Marcotte,E.M. (2004) A probabilistic functional network of yeast genes. *Science*, **306**, 1555–1558.
30. Lee,I., Li,Z. and Marcotte,E.M. (2007) An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. *PLoS ONE*, **2**, e988.
31. Nariai,N., Kolaczyk,E.D. and Kasif,S. (2007) Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*, **2**, e337.
32. Troyanskaya,O.G., Dolinski,K., Owen,A.B., Altman,R.B. and Botstein,D. (2003) A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl Acad. Sci. USA*, **100**, 8348–8353.
33. Friedman,N., Linial,M., Nachman,I. and Pe'er,D. (2000) Using Bayesian networks to analyze expression data. *J. Comput. Biol.*, **7**, 601–620.
34. Teixeira,M.C., Monteiro,P., Jain,P., Tenreiro,S., Fernandes,A.R., Mira,N.P., Alenquer,M., Freitas,A.T., Oliveira,A.L. and Sa-Correia,I. (2006) The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **34**, D446–D451.
35. Monteiro,P.T., Mendes,N.D., Teixeira,M.C., d'Orey,S., Tenreiro,S., Mira,N.P., Pais,H., Francisco,A.P., Carvalho,A.M., Lourenco,A.B. *et al.* (2008) YEASTRACT-DISCOVERER: new tools to improve the analysis of transcriptional regulatory associations in Saccharomyces cerevisiae. *Nucleic Acids Res.*, **36**, D132–D136.
36. Mewes,H.W., Heumann,K., Kaps,A., Mayer,K., Pfeiffer,F., Stocker,S., Frishman,D. and Journals,O. (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **27**, 44–48.
37. Lee,T.I., Rinaldi,N.J., Robert,F., Odom,D.T., Bar-Joseph,Z., Gerber,G.K., Hannett,N.M., Harbison,C.T., Thompson,C.M., Simon,I. *et al.* (2002) Transcriptional regulatory networks in saccharomyces cerevisiae. *Science*, **298**, 799–804.
38. Mani,R., St Onge,R.P., Hartman Iv,J.L., Giaever,G. and Roth,F.P. (2008) Defining genetic interaction. *Proc. Natl Acad. Sci. USA*, **105**, 3461–3466.
39. Luscombe,N.M., Babu,M.M., Yu,H., Snyder,M., Teichmann,S.A. and Gerstein,M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.
40. Bowers,P., Pellegrini,M., Thompson,M., Fierro,J., Yeates,T. and Eisenberg,D. (2004) Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol.*, **5**, R35.
41. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J.B., Reynolds,D.B. and Yoo,J. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
42. Horak,C.E., Luscombe,N.M., Qian,J., Bertone,P., Piccirrillo,S., Gerstein,M. and Snyder,M. (2002) Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev.*, **16**, 3017–3033.
43. Workman,C.T., Mak,H.C., McCuine,S., Tagne,J.-B., Agarwal,M., Ozier,O., Begley,T.J., Samson,L.D. and Ideker,T. (2006) A systems approach to mapping DNA damage response pathways. *Science*, **312**, 1054–1059.
44. Borneman,A.R., Zhang,Z.D., Rozowsky,J., Seringhaus,M.R., Gerstein,M. and Snyder,M. (2007) Transcription factor binding site identification in yeast: a comparison of high-density oligonucleotide and PCR-based microarray platforms. *Funct. Integrative Genomics*, **7**, 335–345.
45. Nguyen,D.H. and D'Haeseleer,P. (2006) Deciphering principles of transcription regulation in eukaryotic genomes. *Mol. Sys. Biol.*, **2**, 0012.
46. Yu,H., Luscombe,N.M., Qian,J. and Gerstein,M. (2003) Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet.*, **19**, 422–427.
47. Ge,H., Liu,Z., Church,G.M. and Vidal,M. (2001) Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat. Genet.*, **29**, 482–486.
48. Lee,H., Tu,Z., Deng,M., Sun,F. and Chen,T. (2006) Diffusion kernel-based logistic regression models for protein function prediction. *OMICS J. Integ. Biol.*, **10**, 40–55.
49. Stark,C., Breitkreutz,B.-J., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
50. Wu,H., Su,Z., Mao,F., Olman,V. and Xu,Y. (2005) Prediction of functional modules based on comparative genome analysis and Gene Ontology application. *Nucleic Acids Res.*, **33**, 2822–2837.
51. Chen,Y. and Xu,D. (2004) Global protein function annotation through mining genome-scale data in yeast Saccharomyces cerevisiae. *Nucleic Acids Res.*, **32**, 6414–6424.
52. Friedman,N., Geiger,D. and Goldszmidt,M. (1997) Bayesian network classifiers. *Machine Learning*, **29**, 131–163.
53. Friedman,N. (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.
54. Demsar,J., Zupan,B. and Leban,G. (2004) Orange: from experimental machine learning to interactive data mining. *Lect. Notes Comput. Sci.*, **3202**, 537–539.
55. Lu,L.J., Xia,Y., Paccanaro,A., Yu,H. and Gerstein,M. (2005) Assessing the limits of genomic data integration for predicting protein networks. *Genome Res.*, **15**, 945–953.
56. Bean,J.M., Siggia,E.D. and Cross,F.R. (2005) High functional overlap between mluI cell-cycle box binding factor and Swi4/6 cell-cycle box binding factor in the G1/S transcriptional program in Saccharomyces cerevisiae. *Genetics*, **171**, 49–61.

57. Iyer,V.R., Horak,C.E., Scafe,C.S., Botstein,D., Snyder,M. and Brown,P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
58. Lotito,L., Russo,A., Bueno,S., Chillemi,G., Fogli,M.V. and Capranico,G. (2009) A specific transcriptional response of yeast cells to camptothecin dependent on the Swi4 and Mbp1 factors. *Eur. J. Pharmac.*, **603**, 29–36.
59. Kumar,R., Reynolds,D.M., Shevchenko,A., Shevchenko,A., Goldstone,S.D. and Dalton,S. (2000) Forkhead transcription factors, Fkh1p and Fkh2p, collaborate with Mcm1p to control transcription required for M-phase. *Curr. Biol.*, **10**, 896–906.
60. Ercan,S., Reese,J.C., Workman,J.L. and Simpson,R.T. (2005) Yeast recombination enhancer is stimulated by transcription activation. *Mol. Cell Biol.*, **25**, 7976–7987.
61. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.L., Millar,A., Taylor,P., Bennett,K. and Boutilier,K. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.
62. Chang,C., Gonzalez,F., Rothermel,B., Sun,L., Johnston,S.A. and Kodadek,T. (2001) The Gal4 activation domain binds Sug2 protein, a proteasome component, in vivo and in vitro. *J. Biol. Chem.*, **276**, 30956–30963.
63. Walther,K. and Schuller,H.J. (2001) Adr1 and Cat8 synergistically activate the glucose-regulated alcohol dehydrogenase gene ADH2 of the yeast Saccharomyces cerevisiae. *Microbiology*, **147**, 2037–2044.
64. Knijnenburg,T.A., de Winde,J.H., Daran,J.M., Daran-Lapujade,P., Pronk,J.T., Reinders,M.J.T. and Wessels,L.F. (2007) Exploiting combinatorial cultivation conditions to infer transcriptional regulation. *BMC Genomics*, **8**, 25.
65. Hermsen,R., Tans,S., ten Wolde,P.R. and Rhodius,V. (2006) Transcriptional regulation by competing transcription factor modules. *PLoS Comput. Biol.*, **2**, e164.
66. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.
67. Yu,H., Braun,P., Yildirim,M.A., Lemmens,I., Venkatesan,K., Sahalie,J., Hirozane-Kishikawa,T., Gebreab,F., Li,N. and Simonis,N. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.
68. Wang,Y., Joshi,T., Zhang,X.-S., Xu,D. and Chen,L. (2006) Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, **22**, 2413–2420.
69. Wang,R.S., Wang,Y., Zhang,X.S. and Chen,L. (2007) Inferring transcriptional regulatory networks from high-throughput data. *Bioinformatics*, **23**, 3056–3064.
70. Bonneau,R., Reiss,D., Shannon,P., Facciotti,M., Hood,L., Baliga,N. and Thorsson,V. (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. *Genome Biol.*, **7**, R36.
71. Zhu,Z., Shendure,J. and Church,G.M. (2005) Discovering functional transcription-factor combinations in the human cell cycle. *Genome Res.*, **15**, 848–855.
72. Yu,X., Lin,J., Zack,D.J. and Qian,J. (2006) Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues. *Nucleic Acids Res.*, **34**, 4925–4936.
73. Hobert,O. (2004) Common logic of transcription factor and microRNA action. *Trends Biochem. Sci.*, **29**, 462–468.
74. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C. and Stoffel,M. (2005) Combinatorial microRNA target predictions. *Nature Genet.*, **37**, 495–500.
75. Breitkreutz,B.J., Stark,C. and Tyers,M. (2003) Osprey: a network visualization system. *Genome Biol.*, **4**, R22.